图书情报工作
第 62 卷 第 12 期 2018 年 6 月

基于 PLoS 开放获取数据的单篇论文网络浏览量累积规律的数理统计及分析

王真 马建华
中国科学院大学 北京 100049 中国科学院文献情报中心 北京 100190

摘要：［**目的/意义**］用数理统计的方法探索 PLoS 平台开放获取的学术论文在网络媒体中浏览量的累积规律,丰富对 Altmetrics 指标的研究方法。［**方法/过程**］跟踪记录 *PLoS Biology* 和 *PLoS Medicine* 期刊 2016 年 11 月份发表的 38 篇研究论文的浏览量数据,数据收集截止到 2017 年 10 月 16 日。利用曲线拟合和计算特别节点等方法对所记录数据进行统计学分析及检验,探索总浏览量指标的累积规律。［**结果/结论**］总浏览量累积曲线与对数曲线拟合优度最高,平均为 0.97;计算出累积曲线的特别节点 S(x,y),将总浏览量累积过程分为集中浏览期和分散浏览期两个阶段,其中集中浏览期约占总阶段的 10%,而浏览量却超过了全部浏览量的 55%。

关键词：单篇论文 Altmetrics 浏览量 曲线拟合 特别节点 累积规律
分类号：G931.1
DOI:10.13266/j.issn.0252－3116.2018.12.010

在科学研究开放化、交互化的大环境下,越来越多的研究人员在科学研究和学术交流过程中应用社会网络工具,如维基百科、博客、社会化书签、微博等,这些社会网络行为产生了丰富的在线活动“印迹”,为计量学研究发展提供了新兴的多元化指标——Altmetrics[1]。Altmetrics 最初提出就是利用学术社交平台上网络数据来弥补传统计量学中基于引用的评价指标的不足,例如运用文献管理平台 CiteULike、Mendeley 中的文献阅读量和推荐量以及开放同行评审 F1000 中的推荐量作为评估文献被利用的情况。Altmetrics 研究之初,Altmetrics 指标与传统引用指标之间的相关性问题是衡量 Altmetrics 指标评价文献价值合理性的标准[2]。目前对 Altmetrics 的研究集中在各指标与传统引用指标之间的相关性研究,以及探讨如何以 Altmetrics 为基础建立学术期刊或学术论文的评价指标体系。研究表明,Altmetrics 所包含的指标中,仅少数指标与引文存在中度相关性,其他多数指标与引文并不相关[3]。另有研究指出,Altmetrics 能够从不同维度揭示科研成果的影响力,且影响力本身也可以是多维度的,比如社会影响力与学术影响力[4]。由于 Altmetrics 指标是由多个不同类型、属性和来源的指标“交织”在一起的集合体,是多维度指标,不能简单地、不加区分地混在一起进行研究[5]。对 Altmetrics 单项指标的研究目前多集中于 Mendeley 和 Twitter。其中 M. Thelwall[6] 以 Mendeley 指标为研究对象,探究为何文献的传统引用指标数量多数情况下与 Mendeley 使用情况不相符,研究发现科研论文内容涉及学科的广泛程度的不同导致文献受众人群数量的差异,进而造成了 Mendeley 使用情况与引文数量差异较大的现象。Q. Ke[7] 等系统地研究了 Twitter 指标,提出利用 Twitter 精确标识科研人员的方法。研究发现,Twitter 在不同学科和不同领域受欢迎程度是不同的,从事社会科学研究的学者使用 Twitter 的频率更高。

本研究认为前人的研究尚有两点不足:①无论是 Altmetrics 综合指标还是单指标的系统研究,研究数据均是某一个特定时间点所采集到的 Altmetrics 数据,缺少从时间变化的角度去探究指标变化趋势的研究;②无论是 Mendeley 还是 Twitter,对单篇论文的覆盖率都无法达到 100%,即并不是所有的论文都会被 Mendeley 用户使用或被 Twitter 用户传播,因此在一定程度上缩小了研究范围。

为了弥补前人研究的不足,本研究利用跟踪统计

作者简介：王真(ORCID:0000－0002－2212－4536),硕士研究生;马建华(ORCID:0000－0002－7945－9150),研究馆员,博士,硕士生导师,通讯作者,E-mail:majh@mail.las.ac.cn。

收稿日期:2017－12－25 修回日期:2018－02－06 本文起止页码:72－83 本文责任编辑:王传清

数据的方法，基于 PLoS article-level metrics 的开放数据，以 *PLoS Biology* 和 *PLoS Medicine* 两种期刊的论文为研究对象，跟踪收集了这些论文从 2016 年 11 月份发表后到 2017 年 10 月的 Altmetrics 数据。从时间变化的角度去探索 Altmetrics 指标变化的趋势。出于对单篇论文的覆盖率以及时效性的考虑，本研究首先选择 Altmetrics 中的总浏览量(View)指标进行重点分析。通过数学方法，发现并归纳出 *PLoS Biology* 和 *PLoS Medicine* 两种期刊论文的网络浏览量的变化和累积规律，为期刊出版者及科研管理者提供参考和借鉴。

1 数据来源

以 PLoS 开放平台上 *PLoS Biology* 和 *PLoS Medicine* 两种期刊发表的论文为研究对象，系统跟踪和记录这两个期刊论文 Altmetrics 指标中的浏览量数据。选择这两种期刊的主要原因有：①*PLoS Biology* 和 *PLoS Medicine* 均属于 PLoS 开放获取平台上的网络期刊，其 Altmetrics 数据便于跟踪并且累积较快，可信程度更高；②这两种期刊分别是生物学和医学领域的高影响力期刊，因此认为其数据量较大，更具有统计学意义。

数据收集工作持续时间约为一年，自 2016 年 10 月底开始，截止日期为 2017 年 10 月 16 日。其间，每天浏览 *PLoS Biology* 和 *PLoS Medicine* 的网站，记录和更新研究论文的浏览量数据。PLoS 网站的 Altmetrics 数据每天更新，这样就为本研究的数据记录提供了基础；但考虑到数据收集的工作量以及浏览量数据在论文发表初期的快速积累特点，本研究在论文发表初期是每天记录数据，随着发表时间的增加降低数据记录的频率，在论文发表 3 个月后记录的频率为每月记录一次。经过近一年的数据跟踪，共收集了 2016 年 11 月份发表在 *PLoS Biology* 上的 22 篇研究论文和发表在 *PLoS Medicine* 上的 16 篇共计 38 篇研究论文近一年的浏览量数据。通过 PLoS 官方对浏览量的分类显示研究[8]，文献总浏览量是各种类型浏览量的合计，见表 1。总浏览量由 5 部分数据构成，分别来自于 PLoS 和 PMC (PubMed Central)两个网站中的 HTML 浏览量、PDF 下载量和 XML 浏览量的总和，其中 PMC 不提供 XML 浏览。

表 1 总浏览量指标包含的各类分指标

	网页浏览量	PDF 下载量	XML 下载量	总计
PLoS	PLoS HTML 页面浏览	PLoS PDF 下载	PLoS XML 下载	PLoS 平台总计
PMC	PMC HTML 页面浏览	PMC PDF 下载		PMC 平台总计
总计	HTML 页面总浏览量	PDF 总下载量	XML 总下载量	总浏览量

2 总浏览量累积趋势及拟合曲线研究

为探究总浏览量的累积规律，将观测值用平滑的曲线连接起来，绘制出论文总浏览量的时序趋势图，见图 1。

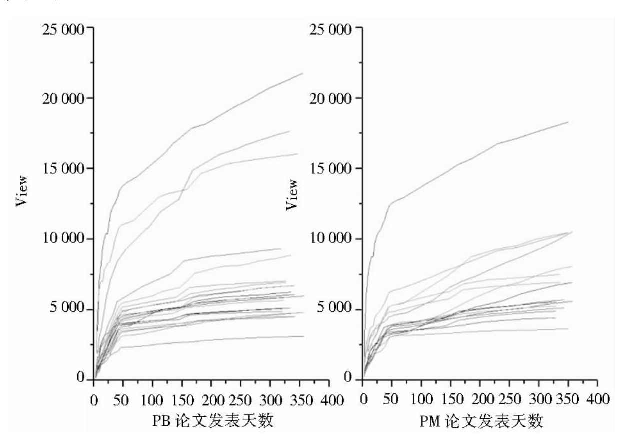


图 1 单篇论文总浏览量随论文发表天数增加而累积的曲线

图 1 中的两个折线图集分别为 *PLoS Biology* 和 *PLoS Medicine* 期刊单篇论文的总浏览量随论文发表天数增加而累积的曲线。很显然，总浏览量累积曲线呈规律性增长：论文发表初期，浏览量急速增加；随着发表天数的增加，曲线开始呈现趋缓的发展态势。这表明，论文一经发表就得到了大量浏览或阅读，而随着发表时间的增加，被浏览的热度有所降低，而到了一定阶段后，浏览量增加的幅度就变得更小。这可以说明研究人员在浏览期刊网站时更倾向于点击最新发表的论文，因而使论文的总浏览量在发表初期快速积累。

2.1 总浏览量累积过程的曲线估计

为进一步探讨总浏览量的变化规律，对总浏览量的观测值进行曲线拟合。通过对实际浏览量的趋势图观察可知，论文发表初期，总浏览量快速累积；但论文发表时间达到一定阶段后，单位时间内浏览量的增量随时间而变小。从函数的角度来看，由于横坐标代表天数，纵坐标代表总浏览量，因此只有第一象限内的函数图像是有意义的。由于浏览量是累积值，随时间的延长而增大，因此拟合函数图像在第一象限内应是增函数，且随 x 的增大曲线的斜率逐渐变小。上述函数曲线的特征与对数曲线相似。

此外，还有一种情况需要考虑，就是个别含有重大发现或内容极具争议性的论文，不像一般论文在发表

一段时间后出现平稳累积状况，而是出现爆发式增长的情况，即论文发表一段时间后浏览量的累积速度加快。函数图像中表现为斜率突然增大，在统计学中，经常用 S 型曲线模型来模拟这种经过一段时间后增长速度突然变快的情况。

因此本研究选择对数曲线模型、S 型曲线模型和 Logistic 模型对一年内总浏览量的累积曲线进行曲线拟合。本研究使用 SPSS 工具，SPSS 给出的拟合模型如公式(1)到公式(3)，其中 e 为自然底数，u 为上界值。

对数曲线：$f(x) = b_0 + b_1 * ln\ x$ 公式(1)

S 曲线：$f(x) = e^{b_0 + b_1/x}$ 公式(2)

Logistic 模型：$f(x) = \frac{1}{(\frac{1}{u} + b_0 * b_1^x)}$ 公式(3)

2.2 总浏览量累积曲线拟合结果

利用 SPSS 软件对总浏览数据的观测值与对数曲线、S 型曲线和 Logistic 模型曲线进行曲线拟合并绘制模型图，拟合结果见表 2 – 表 4，模型图见图 2。

在表 2 – 表 4 中，Article ID 为本研究案例中研究论文的唯一标识符，其中以 PB 开头的 ID 表示该论文发表于 *PLoS Biology*，*PM* 则表示 *PLoS Medicine*，PB 和 PM 后所接的数字为 PLoS 网站提供的论文编号；R^2 为所选模型对观测值的拟合优度决定系数，R^2 的值越接近 1 说明所选模型越能精确描述观测值变化的趋势特征；F 值是回归函数的显著性检验；*df1* 与 *df2* 的和表示每篇文章被跟踪记录的次数，即观测值的个数；*sig* 是 F 检验的概率值；b_0 与 b_1 为拟合曲线函数的两个参数。

设截止到 2017 年 10 月 16 日当天所累积的总浏览量为 V，表 2 – 表 4 中论文出现的次序以 V 升序排序，即位于表中第一行的研究论文 PB_1002571 的总浏览累积量最少，最后一行 PB_1002570 总浏览量累积数量最多。

表 2 对数模型的拟合结果

Article ID	V	R^2	F	*df1*	*df2*	*sig*	b_0	b_1
PB_1002571	3 137	0.97	1 050.37	1	37	0.00	–233.30	597.21
PM_1002161	3 613	0.94	479.99	1	30	0.00	–97.12	686.89
PM_1002172	4 429	0.96	390.74	1	18	0.00	–669.53	918.01
PB_2000391	4 500	0.95	492.51	1	28	0.00	–1 401.35	1 055.17
PB_1002576	4 557	0.95	571.44	1	29	0.00	–963.75	998.26
PB_2000206	4 784	0.96	824.34	1	33	0.00	–1 488.83	1 109.37
PB_2000117	4 821	0.99	6 359.83	1	37	0.00	–605.77	913.69
PB_2000127	4 889	0.96	350.72	1	16	0.00	–1 156.16	1 109.04
PM_1002171	4 906	0.98	746.29	1	18	0.00	–1 121.91	1 056.19
PB_1002580	5 106	0.95	500.35	1	24	0.00	–1 586.38	1 205.07
PB_2000504	5 108	0.96	445.28	1	18	0.00	–1 204.40	1 138.18
PM_1002167	5 127	0.97	760.57	1	24	0.00	–948.67	1 072.14
PM_1002178	5 270	0.96	487.29	1	19	0.00	–1 227.36	1 155.30
PM_1002175	5 561	0.98	922.73	1	19	0.00	–880.71	1 126.62
PM_1002159	5 601	0.98	1 919.28	1	32	0.00	–646.50	1 040.55
PB_2000998	5 653	0.97	678.84	1	18	0.00	–1 586.34	1 299.58
PM_1002166	5 696	0.98	953.71	1	24	0.00	–1 402.16	1 223.54
PB_1002578	5 885	0.96	594.73	1	24	0.00	–1 060.70	1 254.23
PB_1002569	5 986	0.99	3 121.27	1	38	0.00	171.67	1 006.28
PB_2000237	6 081	0.97	995.76	1	28	0.00	–2 061.57	1 415.37
PB_1002577	6 259	0.96	648.43	1	24	0.00	–1 710.78	1 410.61
PB_2000733	6 700	0.98	1 178.05	1	24	0.00	–2 397.38	1 582.33
PB_2000638	6 730	0.98	1 415.03	1	28	0.00	–1 073.67	1 364.52
PM_1002149	6 935	0.95	553.60	1	32	0.00	–1 574.24	1 316.88
PB_1002581	6 941	0.98	1 045.45	1	21	0.00	–2 360.04	1 644.91
PM_1002169	6 948	0.98	1 158.23	1	20	0.00	–955.35	1 395.03
PB_1002579	7 035	0.98	1 106.48	1	22	0.00	–2 023.03	1 630.05
PM_1002170	7 499	0.98	934.81	1	20	0.00	–1 514.45	1 597.87

（续表 2）

Article ID	V	R^2	F	$df1$	$df2$	sig	b_0	b_1
PM_1002155	8 088	0.92	375.61	1	31	0.00	－3 181.36	1 725.35
PB_1002575	8 884	0.98	1 424.06	1	24	0.00	－2 460.17	1 920.67
PB_2000598	9 338	0.99	2 864.34	1	17	0.00	－2 500.76	2 085.75
PM_1002164	10 137	0.99	3 070.23	1	30	0.00	－1 249.03	1 897.55
PM_1002160	10 259	0.96	763.24	1	30	0.00	－2 255.22	2 020.42
PM_1002152	10 533	0.91	333.58	1	31	0.00	－2 310.13	1 890.60
PB_1002573	16 043	0.99	2 380.55	1	31	0.00	－594.83	2 888.60
PB_2000225	17 666	0.99	2 217.82	1	24	0.00	－6 418.30	4 070.51
PM_1002158	17 961	0.99	2 095.09	1	30	0.00	256.58	3 049.60
PB_1002570	21 768	0.99	2 707.61	1	41	0.00	－1 727.39	3 908.23

表 3　S 曲线模型的拟合结果

Article ID	V	R^2	F	$df1$	$df2$	sig	b_0	b_1
PB_1002571	3 137	0.97	1 237.55	1	37	0.00	7.97	－11.35
PM_1002161	3 613	0.86	185.69	1	30	0.00	8.02	－5.86
PM_1002172	4 429	0.95	317.98	1	18	0.00	8.25	－8.83
PB_2000391	4 500	0.87	191.97	1	28	0.00	8.15	－10.46
PB_1002576	4 557	0.92	340.62	1	29	0.00	8.24	－9.66
PB_2000206	4 784	0.92	360.89	1	33	0.00	8.23	－11.85
PB_2000117	4 821	0.92	431.19	1	37	0.00	8.29	－11.01
PB_2000127	4 889	0.96	373.10	1	16	0.00	8.45	－12.04
PM_1002171	4 906	0.89	147.77	1	18	0.00	8.26	－9.12
PB_1002580	5 106	0.91	242.55	1	24	0.00	8.45	－14.65
PB_2000504	5 108	0.94	297.32	1	18	0.00	8.40	－10.75
PM_1002167	5 127	0.90	220.00	1	24	0.00	8.38	－10.18
PM_1002178	5 270	0.88	134.99	1	19	0.00	8.34	－9.04
PM_1002175	5 561	0.93	237.25	1	19	0.00	8.42	－8.29
PM_1002159	5 601	0.95	651.39	1	32	0.00	8.48	－12.02
PB_2000998	5 653	0.90	168.91	1	18	0.00	8.45	－10.59
PM_1002166	5 696	0.84	125.82	1	24	0.00	8.40	－10.22
PB_1002578	5 885	0.86	145.99	1	24	0.00	8.47	－7.85
PB_1002569	5 986	0.94	569.65	1	38	0.00	8.57	－8.91
PB_2000237	6 081	0.80	112.99	1	28	0.00	8.39	－10.31
PB_1002577	6 259	0.92	262.38	1	24	0.00	8.66	－15.19
PB_2000733	6 700	0.91	228.88	1	24	0.00	8.67	－15.95
PB_2000638	6 730	0.92	325.96	1	28	0.00	8.60	－8.60
PM_1002149	6 935	0.80	131.50	1	32	0.00	8.49	－11.59
PB_1002581	6 941	0.93	272.98	1	21	0.00	8.69	－13.74
PM_1002169	6 948	0.94	341.60	1	20	0.00	8.77	－12.07
PB_1002579	7 035	0.94	359.52	1	22	0.00	8.74	－12.48
PM_1002170	7 499	0.94	329.58	1	20	0.00	8.85	－13.45
PM_1002155	8 088	0.79	114.74	1	31	0.00	8.53	－14.79
PB_1002575	8 884	0.88	174.13	1	24	0.00	8.92	－14.25
PB_2000598	9 338	0.96	406.02	1	17	0.00	8.96	－10.73
PM_1002164	10 137	0.90	276.61	1	30	0.00	8.92	－7.19
PM_1002160	10 259	0.82	134.48	1	30	0.00	8.79	－7.29
PM_1002152	10 533	0.76	96.76	1	31	0.00	8.81	－10.99
PB_1002573	16 043	0.91	300.56	1	31	0.00	9.50	－7.57
PB_2000225	17 666	0.91	228.88	1	24	0.00	9.54	－14.24
PM_1002158	17 961	0.94	451.87	1	30	0.00	9.60	－6.29
PB_1002570	21 768	0.94	694.43	1	41	0.00	9.79	－9.80

表 4 Logistic 模型拟合结果

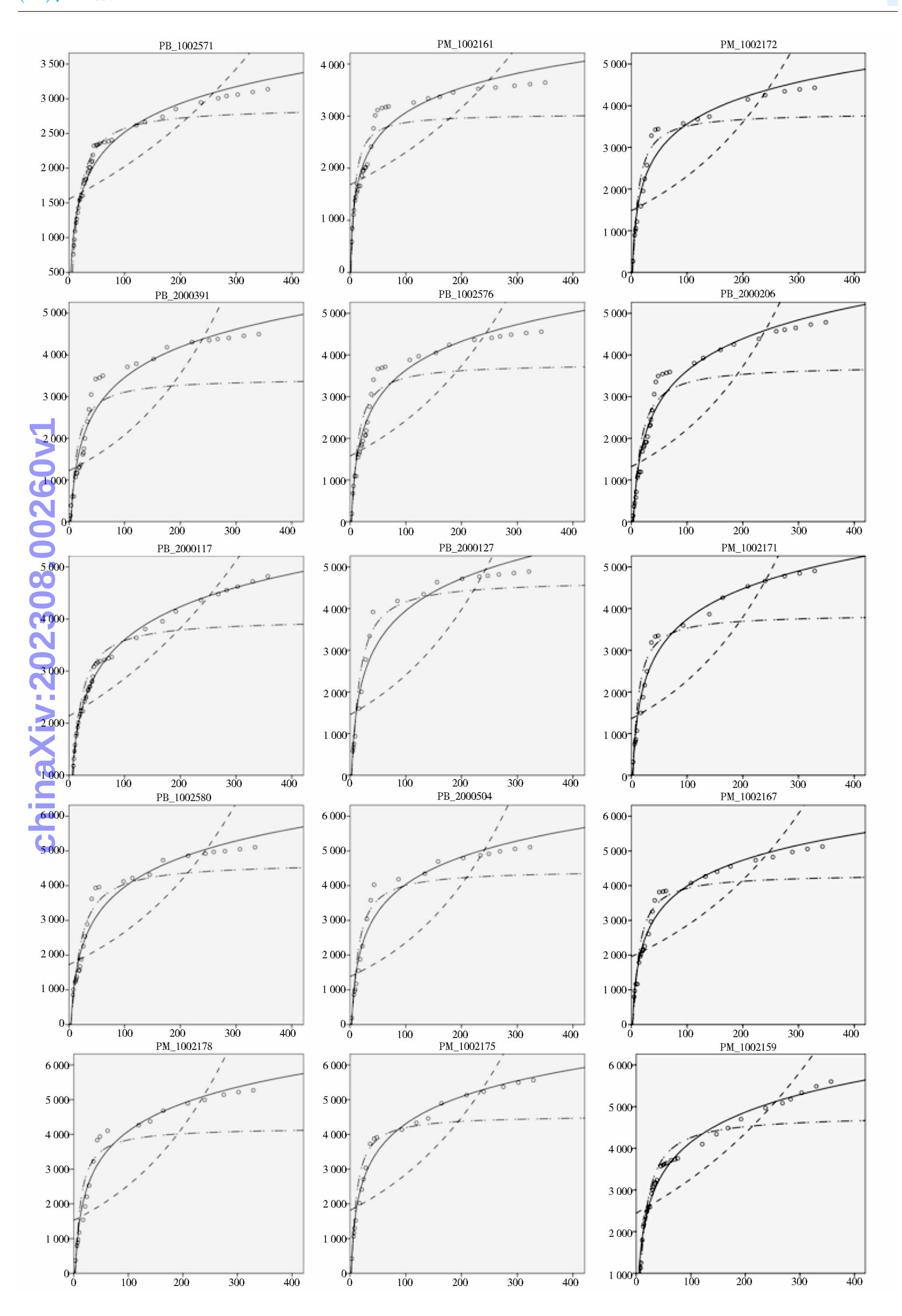
Article ID	V	R^2	F	df1	df2	sig	b_0	b_1
PB_1002571	3 137	0.53	41.76	1	37	0.00	0.00	1.00
PM_1002161	3 613	0.47	26.59	1	30	0.00	0.00	1.00
PM_1002172	4 429	0.44	14.23	1	18	0.00	0.00	1.00
PB_2000391	4 500	0.45	23.07	1	28	0.00	0.00	0.99
PB_1002576	4 557	0.41	20.49	1	29	0.00	0.00	1.00
PB_2000206	4 784	0.41	22.80	1	33	0.00	0.00	0.99
PB_2000117	4 821	0.65	69.27	1	37	0.00	0.00	1.00
PB_2000127	4 889	0.55	19.73	1	16	0.00	0.00	0.99
PM_1002171	4 906	0.53	20.18	1	18	0.00	0.00	0.99
PB_1002580	5 106	0.62	38.40	1	24	0.00	0.00	1.00
PB_2000504	5 108	0.48	16.85	1	18	0.00	0.00	0.99
PM_1002167	5 127	0.54	27.62	1	24	0.00	0.00	1.00
PM_1002178	5 270	0.51	19.99	1	19	0.00	0.00	0.99
PM_1002175	5 561	0.51	19.63	1	19	0.00	0.00	1.00
PM_1002159	5 601	0.61	51.06	1	32	0.00	0.00	1.00
PB_2000998	5 653	0.54	21.04	1	18	0.00	0.00	0.99
PM_1002166	5 696	0.61	37.01	1	24	0.00	0.00	1.00
PB_1002578	5 885	0.54	28.39	1	24	0.00	0.00	1.00
PB_1002569	5 986	0.62	63.05	1	38	0.00	0.00	1.00
PB_2000237	6 081	0.61	43.24	1	28	0.00	0.00	1.00
PB_1002577	6 259	0.67	48.13	1	24	0.00	0.00	1.00
PB_2000733	6 700	0.70	55.74	1	24	0.00	0.00	1.00
PB_2000638	6 730	0.43	20.73	1	28	0.00	0.00	1.00
PM_1002149	6 935	0.84	169.16	1	32	0.00	0.00	1.00
PB_1002581	6 941	0.60	32.07	1	21	0.00	0.00	0.99
PM_1002169	6 948	0.68	41.85	1	20	0.00	0.00	1.00
PB_1002579	7 035	0.59	31.18	1	22	0.00	0.00	0.99
PM_1002170	7 499	0.65	37.73	1	20	0.00	0.00	1.00
PM_1002155	8 088	0.84	160.60	1	31	0.00	0.00	1.00
PB_1002575	8 884	0.77	79.70	1	24	0.00	0.00	1.00
PB_2000598	9 338	0.55	20.62	1	17	0.00	0.00	0.99
PM_1002164	10 137	0.52	31.97	1	30	0.00	0.00	1.00
PM_1002160	10 259	0.64	54.27	1	30	0.00	0.00	1.00
PM_1002152	10 533	0.87	202.97	1	31	0.00	0.00	1.00
PB_1002573	16 043	0.52	33.93	1	31	0.00	0.00	1.00
PB_2000225	17 666	0.71	59.61	1	24	0.00	0.00	0.99
PM_1002158	17 961	0.43	22.86	1	30	0.00	0.00	1.00
PB_1002570	21 768	0.55	49.15	1	41	0.00	0.00	1.00

图 2 中每一张小图为一篇研究论文的曲线估计模型图，总计 38 张。图中圆圈为观测值，实线为拟合对数函数曲线，虚线为 Logistic 模型拟合函数曲线，虚线与点组合的线为 S 曲线拟合函数曲线。

观察模型图发现，对数函数的拟合曲线与观测值的重叠率最大，S 型曲线次之，Logistic 模型曲线观测值几乎没有重叠。虽然以升序的方式对模型图进行了排列，但图形的趋势和拟合优度并没有明显的规律性变化。

2.3 曲线拟合优度决定参数 R^2 的比较

本研究观测的论文样本数量为 38 个，每种曲线拟合产生 38 个 R^2 值，求出每种曲线拟合 R2 的描述性统计量见表 5。R^2 值越接近 1 拟合优度越高，拟合结果显示，对数曲线的拟合优度最高，平均值为 0.97，且对数曲线的 R^2 的极差、方差、平均差和标准差都是最小的，说明论文受关注度的大小不会影响对数曲线的拟合优度。


PB_1002571
PM_1002161
PM_1002172
PB_2000391
PB_1002576
PB_2000206
PB_2000117
PB_2000127
PM_1002171
PB_1002580
PB_2000504
PM_1002167
PM_1002178
PM_1002175
PM_1002159

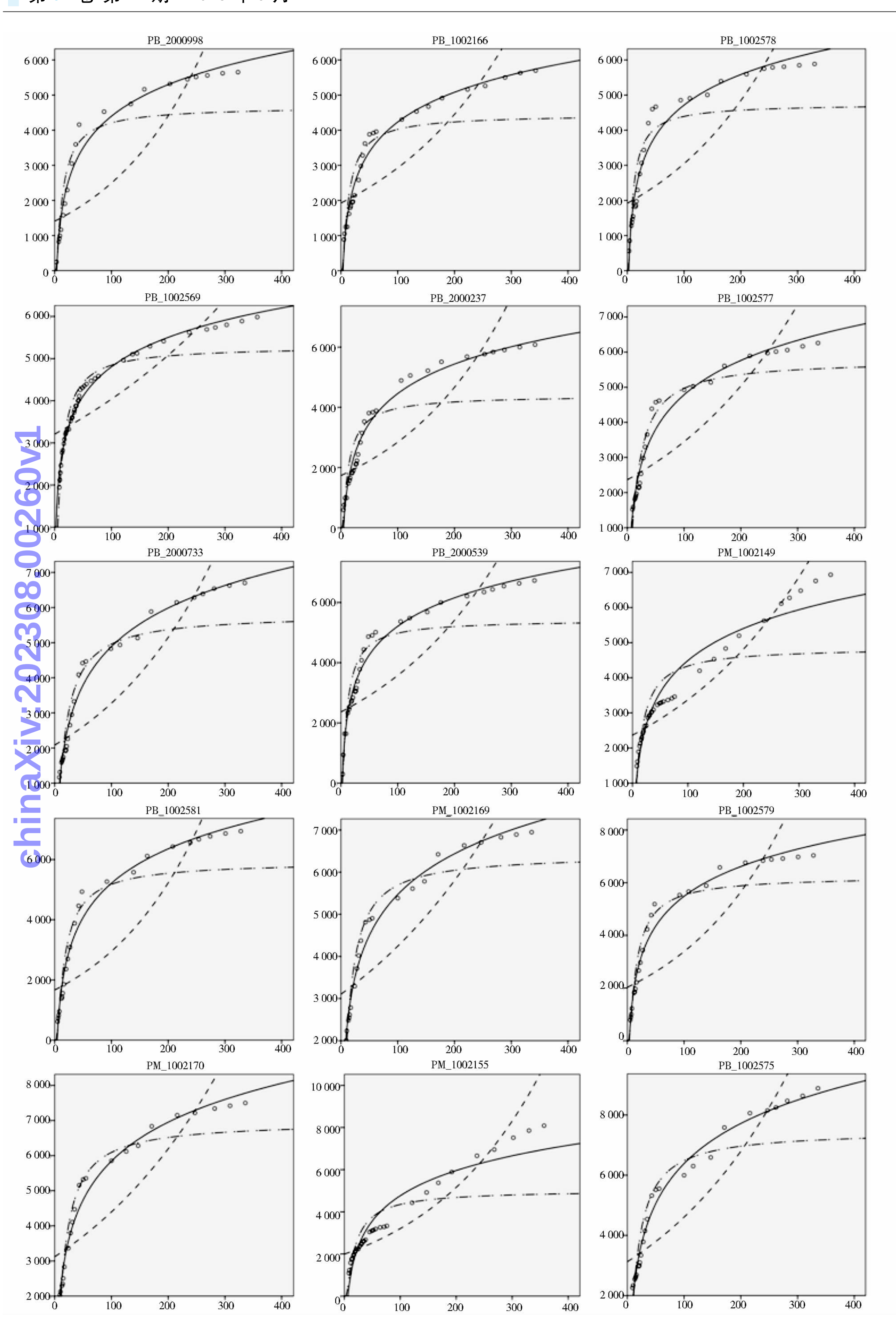

PB_2000998
PB_1002166
PB_1002578
PB_1002569
PB_2000237
PB_1002577
PB_2000733
PB_2000539
PM_1002149
PB_1002581
PM_1002169
PB_1002579
PM_1002170
PM_1002155
PB_1002575

图 2　38 篇案例研究论文总浏览量数据与对数曲线、S 型曲线和 Logistic 曲线的拟合模型

表 5　3 种曲线拟合参数 R2 的描述统计值

拟合模型	有效数据	极差	最小值	最大值	平均值	标准差	方差
对数模型 R^2	38	0.08	0.92	0.99	0.97	0.02	0.00
S 曲线模型 R^2	38	0.21	0.76	0.97	0.90	0.05	0.00
logistic 模型 R^2	38	0.46	0.41	0.87	0.58	0.12	0.01

S 曲线的拟合优度也相对较高,S 曲线的 R^2 值在 0.75 – 0.97 之间波动。

为探究 S 曲线和对数曲线与受关注度不同的论文浏览量累积曲线拟合的优劣,将 R2 值按 View 升序排列并给于 1,2,3,……,38 的序号,绘制出以序号为横坐标,以对数曲线和 S 曲线拟合优度决定参数 R^2 的带点折线图,见图 3。

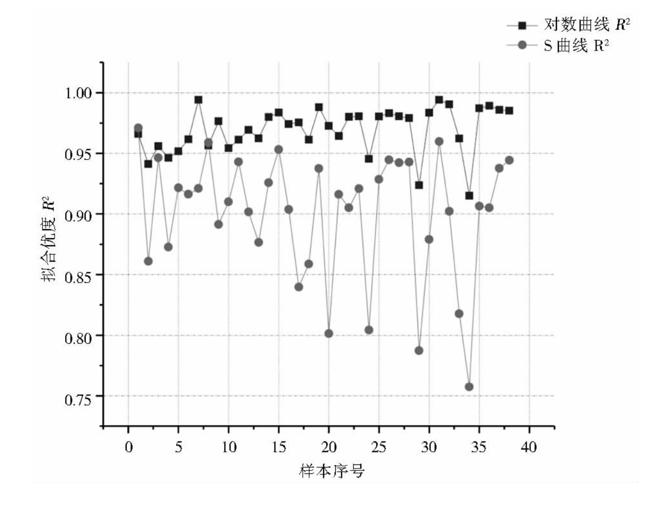


图 3　对数曲线与 S 曲线拟合优度决定参数 R2 折线图

观察图 3 发现,对数曲线的 R^2 折线图总体在 S 曲线的 R2 折线图的上方,S 曲线的 R^2 值在折线图的左端更接近 1,并且 1 号论文 S 曲线的 R2 超过了对数曲线的 R^2。除了 1 号和 8 号论文外,其余论文的浏览量的累积曲线都与对数曲线的拟合优度更高。

通过对总浏览量的累计趋势曲线的研究发现，研究人员在网络中更倾向于点击最新发表的论文，因此总浏览量在论文发表初期便能够快速累积，其拟合曲线和对数曲线和 S 型曲线的前期都是比较一致的。而在中期阶段，S 型曲线和对数曲线有明显的不同，S 曲线在达到饱和值后曲线会趋向平稳，而对数曲线的增长速度虽然也在减小但却并不会出现一个几乎不增加的平台区。现实中，研究论文的网络浏览量没有表现出明显的临界值，因此用对数曲线模拟论文的总浏览累积曲线更合理。

受关注度相对较低的研究论文，其累积总浏览量似乎与 S 曲线的拟合优度更高，说明这些论文发表后，初期也获得了研究人员的关注，但是，之后就不再被人浏览和阅读。该现象背后的原因尽管不是很清楚，但存在以下两种可能性：①这些论文的学术价值可能不是特别高，因此其浏览量到达一个类似临界值后便几乎不再累积，近似 S 曲线的成熟期；②这类研究并不是相应领域的热点问题或研究前沿，或者说，这类研究是一个相对窄的或者冷僻的方向，并不会引起更多人的持续关注。

3　集中浏览期特别节点的理论探寻及价值

3.1　设置特别节点 $S(x,y)$ 的意义

通过上述研究可以发现，论文的总浏览量变化非常符合对数曲线，即总浏览量在经历了快速增长期之后，其单位时间内的增长速率开始变小，总浏览量呈现稳步、小幅度增加。本研究把总浏览量快速累积的阶段称为"集中浏览期"，把之后的平稳阶段称为"分散浏览期"。此两个阶段之间会有一个节点，本研究称之为"总浏览量累积过程的特别节点"，用 $S(x,y)$ 表示。由于对数函数本身为增函数，在数学意义上没有斜率变化的分界点即拐点，因此 $S(x,y)$ 是人为定义的特别节点。

一篇研究论文浏览量的特别节点 $S(x,y)$ 表示该论文发表 x 天后，单位时间内增加的浏览量开始逐渐稳定，这些天的总浏览量为 y。由于每篇论文的影响力不同，其受到的关注程度也不同，因此每篇论文的 $S(x,y)$ 的位置也是不同的。只有标准化后的特别节点 $S(x,y)$ 才能在不同论文之间进行比较，进而可在一定程度上揭示一篇论文发表初期的影响力与受关注度。

3.2　特别节点 $S(x,y)$ 的设置方法

由于 $S(x,y)$ 本身是人为定义的对数函数曲线上的特别节点，因此本研究认为其标准化的计算方法不是唯一的。不同平台，不同开放类型的期刊的集中浏览期是不同的，本研究只针对 PLoS 开放平台上的总浏览量累积曲线提供一种可能的 $S(x,y)$ 的标准化计算方法。

已知用对数函数拟合总浏览量累积曲线的表达式为：

$$f(x) = b_0 + b_1 * ln\ x \qquad 公式(4)$$

为寻找单位时间内函数增量出现变化的点，即斜率变化趋势的特殊点，求出拟合函数 $f(x)$ 的一阶导数，表达式为：

$$f(x)' = b_1/x \qquad 公式(5)$$

$f(x)'$ 是反比例函数，由反比例函数的性质可知当 $x>0$ 时 $f(x)'$ 是减函数，为了寻找原函数 $f(x)$ 斜率变化的特殊点即 $f(x)'$ 曲线图的特殊点，画出反比例函数 $y=1/x$ 在第一象限内的图像见图 4 中曲线，其中，$x \in (0,10]$：

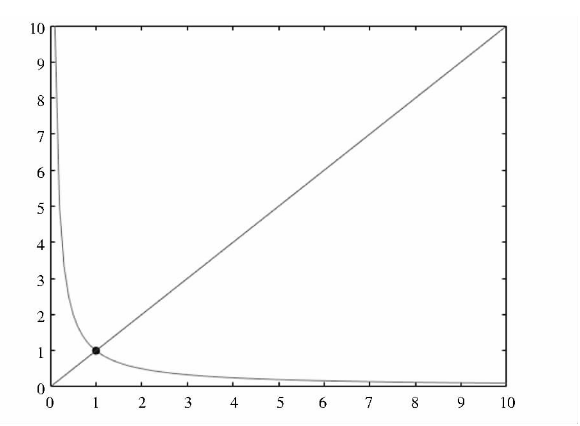


图 4　$y=1/x$ 与 $y=x$ 两函数当 $x \in (0,10]$ 时的函数图像

总浏览量累积曲线的拟合原函数 $f(x)$ 在第一象限内为增函数，因此在本研究中 $f(x)'$ 只在第一象限内的数据是有意义的。$f(x)'$ 在第一象限内为减函数，其单位时间的减量当到达距离原点(0,0)的平面距离最近的点之后单位时间内的减量变化逐渐平稳，即如图中直线与曲线的交点。$f(x)'$ 曲线随 x 增大而减小的幅度从这一点开始明显变缓。$f(x)'$ 的这一变化，对应 $f(x)$ 曲线从这一点开始增长速度逐渐平稳，逐渐平稳且趋于 0 的斜率将使 $f(x)$ 看上去更像一条斜率为常数的直线。

通过数学推算可以确定：反比例函数距离原点(0,0)平面距离最近的点即与 $y=x$ 直线的交点，即图 4 中的曲线与直线的交点 $(\sqrt{b_1}, \sqrt{b_1})$。

因此原函数 $f(x) = b_0 + b_1 * ln\ x$ 的特别节点 $S(x,$

y)为($\sqrt{b_1}$，$f(\sqrt{b_1})$)，可计算本研究案例中的 38 篇论文的总浏览量累积曲线特别节点 $S(x,y)$见表 6。设截至 2017 年 10 月 16 日，文章共发表了 D 天，累积总浏览量为 V，表 6 按 V 升序排列。

表 6　样本论文总浏览量累积曲线在特别节点的数据计算结果

Article ID	V	D	b_0	b_1	x	y	P_x	P_y
PB_1002571	3 137	357	－233.30	597.21	24.44	1 675.45	6.85%	53.41%
PM_1002161	3 613	323	－97.12	686.89	26.21	2 146.32	8.11%	59.41%
PM_1002172	4 429	329	－669.53	918.01	30.30	2 461.89	9.21%	55.59%
PB_2000391	4 500	341	－1 401.35	1 055.17	32.48	2 271.43	9.53%	50.48%
PB_1002576	4 557	343	－963.75	998.26	31.60	2 483.25	9.21%	54.49%
PB_2000206	4 784	349	－1 488.83	1 109.37	33.31	2 400.36	9.54%	50.17%
PB_2000117	4 821	357	－605.77	913.69	30.23	2 508.78	8.47%	52.04%
PB_2000127	4 889	321	－1 156.16	1 109.04	33.30	2 731.71	10.37%	55.87%
PM_1002171	4 906	329	－1 121.91	1 056.19	32.50	2 554.92	9.88%	52.08%
PB_1002580	5 106	334	－1 586.38	1 205.07	34.71	2 688.16	10.39%	52.65%
PB_2000504	5 108	323	－1 204.40	1 138.18	33.74	2 800.40	10.44%	54.82%
PM_1002167	5 127	343	－948.67	1 072.14	32.74	2 791.70	9.55%	54.45%
PM_1002178	5 270	329	－1 227.36	1 155.30	33.99	2 846.32	10.33%	54.01%
PM_1002175	5 561	329	－880.71	1 126.62	33.57	3 077.67	10.20%	55.34%
PM_1002159	5 601	357	－646.50	1 040.55	32.26	2 968.11	9.04%	52.99%
PB_2000998	5 653	323	－1 586.34	1 299.58	36.05	3 072.53	11.16%	54.35%
PM_1002166	5 696	343	－1 402.16	1 223.54	34.98	2 947.21	10.20%	51.74%
PB_1002578	5 885	330	－1 060.70	1 254.23	35.42	3 413.33	10.73%	58.00%
PB_1002569	5 986	357	171.67	1 006.28	31.72	3 650.39	8.89%	60.98%
PB_2000237	6 081	342	－2 061.57	1 415.37	37.62	3 072.81	11.00%	50.53%
PB_1002577	6 259	336	－1 710.78	1 410.61	37.56	3 403.93	11.18%	54.38%
PB_2000733	6 700	335	－2 397.38	1 582.33	39.78	3 430.86	11.87%	51.21%
PB_2000638	6 730	341	－1 073.67	1 364.52	36.94	3 851.27	10.83%	57.23%
PM_1002149	6 935	357	－1 574.24	1 316.88	36.29	3 155.36	10.16%	45.50%
PB_1002581	6 941	328	－2 360.04	1 644.91	40.56	3 730.61	12.37%	53.75%
PM_1002169	6 948	336	－955.35	1 395.03	37.35	4 095.12	11.12%	58.94%
PB_1002579	7 035	328	－2 023.03	1 630.05	40.37	4 005.20	12.31%	56.93%
PM_1002170	7 499	336	－1 514.45	1 597.87	39.97	4 378.85	11.90%	58.39%
PM_1002155	8 088	357	－3 181.36	1 725.35	41.54	3 248.33	11.64%	40.16%
PB_1002575	8 884	336	－2 460.17	1 920.67	43.83	4 800.39	13.04%	54.03%
PB_2000598	9 338	320	－2 500.76	2 085.75	45.67	5 469.80	14.27%	58.58%
PM_1002164	10 137	323	－1 249.03	1 897.55	43.56	5 912.61	13.49%	58.33%
PM_1002160	10 259	323	－2 255.22	2 020.42	44.95	5 433.53	13.92%	52.96%
PM_1002152	10 533	357	－2 310.13	1 890.60	43.48	4 821.83	12.18%	45.78%
PB_1002573	16 043	347	－594.83	2 888.60	53.75	10 914.13	15.49%	68.03%
PB_2000225	17 666	333	－6 418.30	4 070.51	63.80	10 497.76	19.16%	59.42%
PM_1002158	17 961	323	256.58	3 049.60	55.22	12 489.71	17.10%	69.54%
PB_1002570	21 768	355	－1 727.39	3 908.23	62.52	14 434.76	17.61%	66.31%

除计算的标准化特别节点 $S(x,y)$，表 6 中也计算统计了拟合对数函数的两个参数 b_0 和 b_1，以及每篇论文的 x 与 D 的比值 P_x，以及 y 与 V 的比值 P_y。P_x 和 P_y 分别表示论文发表到 S 点这一天，其发表天数和累积的总浏览量占论文发表 D 天之后对应数值的百分比。计算 P_x 和 P_y 的描述统计量如表 7 所示：

表 7　P_x 和 P_y 的描述统计量

	最小值	最大值	平均值	标准偏差
P_x	6.85%	19.16%	11.39%	2.63%
P_y	40.16%	69.54%	55.08%	5.64%

表 7 显示，本研究所求特别节点的效果非常显著，所划分的集中浏览期平均只占总发表天数的11.39%，但集中浏览期累积的总浏览量却达到了全程累积的总浏览量的55.08%。即研究论文在约 40 天的集中浏览期中所获得的浏览量超过了其一年内所能累积浏览量的一半以上。因此，可以初步认为，根据集中浏览期的数据可以基本判断该论文一年的浏览量情况，也就是说可以用较短的时间来预测某期刊论文浏览量在一年中的发展状况。

3.3 特别节点 *S* 的位置及其分布规律

为了更好地观察特别节点 $S(x,y)$ 的位置，利用软件绘制每篇论文总浏览量累计观测值所拟合的对数函数的曲线并用圆点标注出 $S(x,y)$，如图 5 所示：

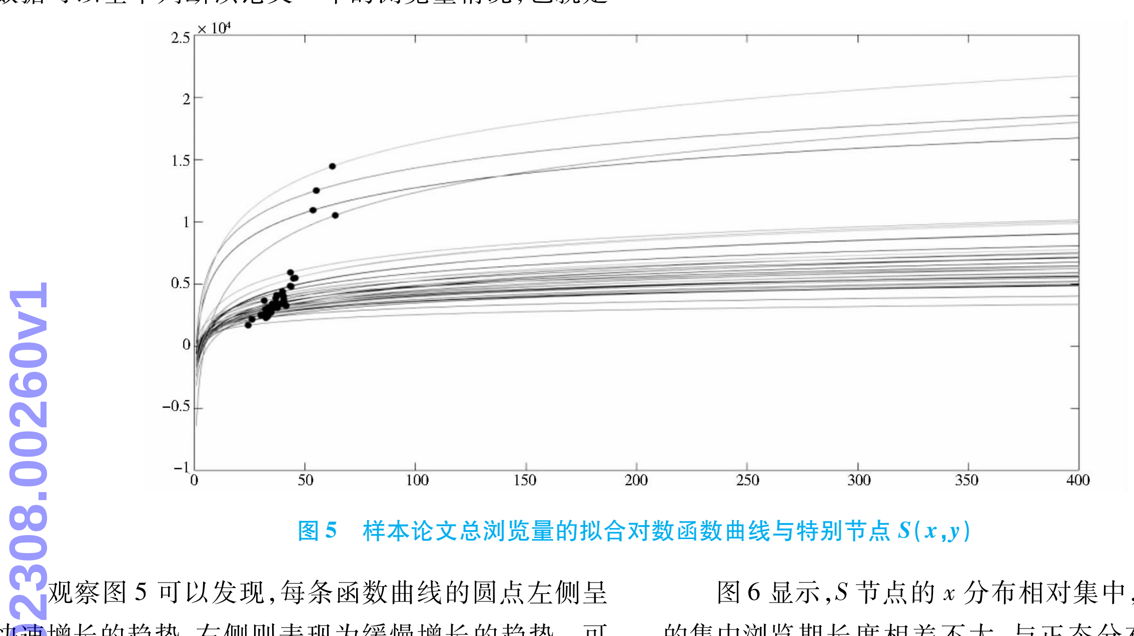


图 5 样本论文总浏览量的拟合对数函数曲线与特别节点 $S(x,y)$

观察图 5 可以发现，每条函数曲线的圆点左侧呈快速增长的趋势，右侧则表现为缓慢增长的趋势。可以直观地看出，特别节点成功地将总浏览量累积曲线分成了集中阅读阶段和分散阅读阶段两个阶段。

除此之外，图 5 显示总浏览量累积量越高，即受关注程度越大的论文，其特别节点 S 的 *x* 值越大。计算出 V 与节点 *S* 的 *x* 值的皮尔逊相关系数为 0.96，表示 *x* 的大小与论文受关注程度呈现出非常显著的正相关关系。说明论文受到的关注程度越高，其集中浏览期持续的时间越长。

为研究集中浏览期的时间长度，即节点 *S* 的 *x* 值的分布规律，作出 x 的频率直方图，如图 6 所示：

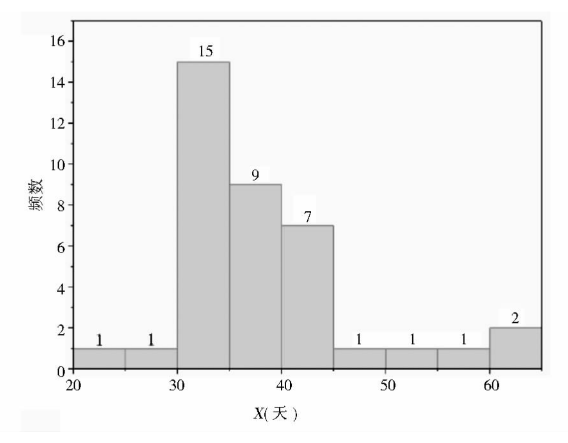


图 6 *S* 点 *x* 的频率直方图

图 6 显示，*S* 节点的 *x* 分布相对集中，即样本论文的集中浏览期长度相差不大，与正态分布相似，利用 SPSS 的单样本 K-S 检验，对 *x* 进行常态性检验，结果见表 11。*S* 节点的 *x* 平均值为 38，即集中浏览期平均为 38 天，符合正态分布。

表 11 节点 S 的横坐标 x 的单样本 Kolmogorov-Smirnov 检验结果

x 的单样本 Kolmogorov-Smirnov 检验		
计数		38
正态参数[a,b]	平均值	38.38
	标准偏差	8.73
最极端差分	绝对	0.17
	正	0.17
	负	-0.12
检验统计		0.17
渐近显著性（双尾）		.010[c]

a. 检验分布是正态分布。
b. 根据数据计算。
c. Lilliefors 显著性校正。

4 结果及讨论

通过对 *PLoS Medicine* 和 *PLoS Biology* 两种期刊 2016 年 11 月发表的 38 篇研究论文（其中 *PLoS Medicine* 发表 16 篇，*PLoS Biology* 发表 22 篇）的浏览量数据的追踪统计，本研究得出下列研究结果。

（1）总浏览量累积曲线呈规律性增长，增长速度先快后慢，说明研究人员在浏览期刊网站时更倾向于

点击最新发表的论文。

(2)对总浏览量的累计曲线进行曲线拟合,发现其与对数函数曲线的拟合优度非常高,其平均值为0.97,且拟合优度不受论文所受关注度的影响。

(3)本研究定义了特别节点 $S(x,y)$ 以及单篇论文的集中浏览期和分散浏览期的意义,并给出了一种可能的特别节点的标准化计算方法。利用本文提出的标准化计算方法所计算出的特别节点效果显著,从而可以利用短暂的集中浏览期所累积的浏览量数据预测其一年内的总浏览量累积情况,及早发现高影响力论文。

(4)通过对集中浏览期分布规律研究发现,样本论文的集中浏览期平均为38天,呈现出正态分布的趋势。并发现论文受到的关注程度越高,其集中浏览期持续的时间越长,皮尔逊相关系数为0.96。

在研究对象方面,相较于目前受学者重视程度较高但覆盖率不高的Twitter和Mendeley等指标,本研究提出应首先深入研究覆盖率最高且累积速度最快的浏览类指标,使更多的被Altmetrics标记的论文存在研究的价值。其次,在研究方法上,本研究弥补了以往人为选择性地下载某一个时间点的Altmetrics数据作为研究数据的单一性,使用跟踪记录的方法收集Altmetrics数据,动态地探索Altmetrics的变化规律,丰富了对Altmetrics的研究方法。

由于研究样本均来自于PLoS旗下的高影响力期刊论文,加之PLoS为开放平台,对论文的访问不受是否订阅的限制,因此,这些论文在发表后很短的时间内就获得了大量的浏览量数据,这些坚实的数据为本研究的结论及可靠性奠定了基础。然而,假如研究对象不是在OA的开放平台,或者说研究论文所在期刊不是领域内的重要期刊,会得出怎样的结论,现在还是不得而知。未来研究将选择不一样特点的期刊论文,进一步发现各种类型论文的传播特征和规律,完善现有的研究结论,对期刊出版提供更科学的理论依据。

参考文献:

[1] 刘菊红,黄凯文. 开放获取资源评价模式的研究:基于单篇论文质量评价[J]. 新世纪图书馆, 2015(5):91-96.

[2] 曹丽江,周毅. 基于元分析的Altmetrics指标与传统引用指标相关性研究[J]. 情报理论与实践, 2016, 39(8):49-53.

[3] PREM J, GROTH P, TARABORELI D. The Altmetrics collection [J]. PLOS ONE, 2012, 7(11): e48753.

[4] 宋丽萍,王建芳,刘芮. 基于主成分分析的科学评价维度研究——以PLoS ONE为例[J]. 图书情报工作, 2014, 58(17): 119-124.

[5] 刘丽敏,王晴. 我国图情领域Altmetrics研究评述及展望[J]. 情报杂志, 2016, 35(4):131-136,124.

[6] KE Q, AHN Y Y, SUGIMOTO C R. A systematic identification and analysis of scientists on Twitter[J]. PLoS ONE, 2016, 12(4): e0175368.

[7] MIKE T. Why do papers have many mendeley readers but few Scopus-indexed citations and vice versa? [J]. Journal of librarianship and information science, 2017, 49(2):144 -151.

[8] Public Library of Science (PLOS)[EB/OL]. [2017-12-04]. http://www.lagotto.io/plos/.

作者贡献说明:

王真:数据收集、统计分析及论文撰写;
马建华:论文框架搭建与论文修改。

Mathematical Statistics and Analysis of Cumulative Rules of Single Paper Web Browsing Based on PLoS Open Access Data

Wang Zhen Ma Jianhua
University of Chinese Academy of Sciences, Beijing 100049
National Science Library, Chinese Academy of Sciences, Beijing 100190

Abstract: [**Purpose/significance**] The purpose of this paper is to research the cumulative rules of PLoS OA articles' views in online media and to enrich the research methods of Altmetrics by mathematical statistics. [**Method/process**] The tracking data for 38 research papers published in November 2016 by *PLoS Biology* and *PLoS Medicine* are collected until October 16, 2017. By using the method of curve fitting and special node calculation, the statistical analysis and test of the recorded data are carried out to explore the cumulative rule of the total page view index. [**Result/conclusion**] The goodness of fit of logarithmic curve to the cumulative curve of total views is the best, with an average of 0.97. The special point $S(x,y)$ of the cumulative curve is calculated, and the cumulative curve of total views is successfully divided into two stages: centralized view period and decentralized view period, of which the centralized view period accounts for about 10% of all stage, while the view volume exceeds 55% of all stage.

Keywords: single article Altmetrics page view curve fitting special point cumulative rules